Calibration and Reliability in Groundwater Modelling (Proceedings of the ModelCARE 99 Conference held at Zürich, Switzerland, September 1999). IAHS Publ. no. 265, 2000.

Guidelines for model calibration and application to flow simulation in the Death Valley regional groundwater system

MARY C. HILL, FRANK A. D'AGNESE &
CLAUDIA C. FAUNT
US Geological Survey 3215 Marine Street, Boulder, Colorado 80303-1066, USA
e-mail: mchill@usgs.gov

Abstract Fourteen guidelines are described which are intended to produce calibrated groundwater models likely to represent the associated real systems more accurately than typically used methods. The 14 guidelines are discussed in the context of the calibration of a regional groundwater flow model of the Death Valley region in the southwestern United States. This groundwater flow system contains two sites of national significance from which the subsurface transport of contaminants could be or is of concern: Yucca Mountain, which is the potential site of the United States high-level nuclear-waste disposal; and the Nevada Test Site, which contains a number of underground nuclear-testing locations. This application of the guidelines demonstrates how they may be used for model calibration and evaluation, and also to direct further model development and data collection.

INTRODUCTION

Lack of data, inaccessibility, scaling issues and the urgent needs of society to make important water-resource decisions make groundwater modelling one of the most challenging model calibration situations in the natural sciences. A number of methods have been suggested for addressing groundwater model calibration; sophisticated statistical techniques and assumptions that are unfamiliar to most groundwater modellers generally characterize these methods. The most fundamental goal of all these methods is to achieve tractable model calibrations that produce meaningful model results. These methods require an understanding of how, and in what circumstances, they can be used most effectively. Hill (1998) describes 14 guidelines for using nonlinear-regression methods for model calibration. These guidelines emphasize the types of situations encountered when calibrating groundwater models. This paper briefly describes these guidelines and demonstrates their use in developing a model of the Death Valley regional groundwater flow system; it underlies the United States' potential high-level nuclear waste repository in the unsaturated zone at Yucca Mountain, and a number of underground nuclear-testing sites on the Nevada Test Site. The possibility of current or potentially contaminated groundwater moving from these sites is of concern. This paper is intended to encourage comments about the methods used to study this important groundwater system.

FOURTEEN GUIDELINES FOR EFFECTIVE MODEL CALIBRATION

There are different opinions about how to calibrate groundwater models, and no single set of ideas is best for all situations. However, considering one complete set of guidelines and associated methods can be advantageous for a number of reasons, including:

- The applicability of the guidelines to some circumstances probably will be generally recognized, despite the wide range of viewpoints about model calibration.
- The guidelines may be used by less experienced modellers to produce calibrated models that are more likely to be accurate than models produced using more traditional calibration methods. Thus, the guidelines can improve the general technical level and likely accuracy of groundwater models.
- The guidelines can be used by more experienced modellers as a basic set of ideas and criteria upon which they can expand.
- The guidelines can provide a set of standard analyses of model fit and sensitivity which groundwater managers and other users of model results can come to expect. These standard analyses can help the managers evaluate the accuracy and reliability of a model, and the appropriateness of the model for their needs.

Natural groundwater systems are so complex that it would be detrimental to impose any set of guidelines too rigorously, but present modelling efforts could be improved substantially through the development, application, and debate of the 14 guidelines (Hill, 1998) summarized in Table 1. These guidelines can be thought of as organized common sense that offers order, perspective, and statistics to the calibration process.

From the perspective of stochastic groundwater inverse methods, the approach presented in the 14 guidelines can be thought of as a strategy to approximate large-scale mean, or effective, values. Stochastic methods generally require that the mean of any spatially distributed quantity, such as hydraulic conductivity, be a constant or a simple function. Unfortunately, geological media often defy these limitations. The methods presented here can be used to test the assumptions needed by most stochastic methods, and, if needed, approximate potentially complex spatial distributions.

Once large-scale variations are established other methods may be used to assess the influence of important variations that may be smaller than the grid scale. To date, methods of determining large-scale variations and methods of characterizing the likely effects of small-scale variations, have been integrated very little (Kitanidis, 1998).

Many aspects of the guidelines presented here are applicable regardless of how a model is calibrated. Some aspects, however, depend on (a) the availability of calculated sensitivities, (b) the use of nonlinear regression to determine parameter values that fit measured values optimally using a least squares criterion, and (c) linear and nonlinear inferential statistics. Sensitivities are the derivatives of simulated values such as flows, hydraulic heads, and concentrations, with respect to defined parameters that represent model inputs such as hydraulic conductivity, areal recharge, and the hydraulic connection of springs to the subsurface. The three listed capabilities are becoming increasingly available with the development of computer programs such as MODFLOWP (Hill, 1992), PEST (Doherty, 1994), and UCODE (Poeter & Hill, 1998). MODFLOWP and UCODE produce the statistics mentioned in the guidelines.

Table 1 Guidelines for effective model calibration.

Guideline		Examples of relevance to the Death Valley regional model (unless indicated, discussed by D'Agnese *et al.*, 1998)
1.	Apply the principle of parsimony (start very simple; build complexity slowly)	– Started with 10 defined parameters: 4 hydraulic conductivity, 2 vertical anisotropy, 1 recharge, 1 evapotranspiration, 1 spring conductance, and 1 pumpage. Added 5 hydraulic conductivity, 3 recharge, 4 spring conductance, and 1 pumpage for 23 final defined parameters. – Used three model layers, the minimum needed to represent expected deeper regional and shallower localized flow paths. This kept forward execution times to about 30 min, allowed thorough analysis of the model and comparison with the regression data, and appeared to be adequate.
2.	Use a broad range of information to constrain the problem	Extensive geological and hydrological data were organized and evaluated using two- and three-dimensional database and visualization methods and results were used to constrain the problem.
3.	Maintain a well-posed, comprehensive regression problem	Composite scaled sensitivities and parameter correlation coefficients were used to measure the information available from the regression data. Regression was used to estimate all parameters for which there was sufficient information.
4.	Include many kinds of data as observations in the regression	Regression data included hydraulic heads and spring flows. No appropriate measurements of regional transport (such as hydrochemistry and environmental isotope analyses) were available for model calibration.
5.	Use prior information carefully	No prior information was used in this version of the model. Omitting prior information promotes understanding of how well the regression data constrain the model. Some parameters were fixed during the regression because of insensitivity, extreme correlation, or unrealistic flow dynamics.
6.	Assign weights which reflect measurement errors	The weights were based on standard deviations and coefficients of variation that reflected expected errors in the measured hydraulic heads and spring flows, respectively. The errors considered included some expected discretization errors.
7.	Encourage convergence by making the model more accurate	Examination of the simulated results and re-evaluation of the geohydrological framework revealed several situations in which the model representation was inaccurate, and these were corrected.
8.	Evaluate model fit	The standard error of the regression, maps of weighted residuals, and graphs of weighted residuals versus weighted simulated values, were the most effective ways of evaluating model fit.
9.	Evaluate optimized parameters	Estimated parameter values were compared to possible ranges. Many of the ranges for parameters were wide. Applying the idea, however, that optimal estimates that fell outside these ranges implied significant inaccuracies in the conceptual model proved to be an effective constraint in model development.
10.	Test alternative models	Many alternative models were considered. For example, inclusion of relatively subtle geological features originally excluded from the geohydrological framework improved model fit.
11.	Evaluate potential new data	Being evaluated but not yet published.
12.	Evaluate the potential for additional estimated parameters	The final model is evaluated in D'Agnese *et al.* (in press) and also is evaluated in this paper.
13.	Use confidence and prediction intervals to indicate parameter and prediction uncertainty	Linear confidence intervals were used to indicate whether estimated parameter values were significantly different and whether they should be combined. Analysis of prediction uncertainty has not been completed.
14.	Formally reconsider the model calibration from the perspective of the desired predictions	Discussed in this paper.

Although model calibration efforts are emphasized in the present paper, many of the methods presented in the guidelines are also useful in other situations in which data and model results are being compared. Barth *et al.* (1999) and Mehl *et al.* (1999) use the methods to evaluate results from controlled intermediate-scale laboratory experiments.

THE DEATH VALLEY REGIONAL GROUNDWATER FLOW SYSTEM

Yucca Mountain in southern Nevada, USA, is being studied as a potential site for a high-level radioactive waste repository. In addition, nearby sites located on the Nevada Test Site have been used for underground testing of nuclear devices. In cooperation with the US Department of Energy, the US Geological Survey is evaluating the geological and hydrological characteristics of the regional groundwater flow system underlying these areas. The Death Valley regional groundwater flow system encompasses nearly 80 000 km^2 and extends from immediately west of Las Vegas, Nevada, to Death Valley National Park, California. Water levels in the region range from more than 1500 m above to 86 m below sea level. The hydrology of the region is the result of both arid climatic conditions and complex geology. Groundwater flow generally can be described as dominated by interbasin flow and may be conceptualized as having two main components: a series of relatively shallow and localized flow paths that are superimposed on deeper regional flow paths. A significant component of the regional groundwater flow is through a thick sequence of Palaeozoic carbonate rock that generally occurs at depth. Regional groundwater flow probably is controlled by extensive and prevalent structural features and abrupt juxtaposition of geological units that result from faulting and fracturing. Water discharges from the system as evapotranspiration by plants, evaporation from playa surfaces, and flow to springs and wells. Water recharges the system mostly as infiltration of precipitation in highlands such as the Spring Mountains and Pahute Mesa. The flow system is hydrogeologically complex and very heterogeneous, with possible local values of hydraulic conductivity ranging over 14 orders of magnitude, and hydraulic gradients ranging from nearly zero to over 2%.

The Death Valley regional groundwater flow system is evaluated in D'Agnese *et al.* (1998) using a fully three-dimensional geoscientific information system and a three-dimensional, steady state, finite difference simulation. The grid has 163 rows, 153 columns, and three layers. The grid cells are oriented north–south and are of uniform size, with side dimensions of 1500 m. The layers span depths below the estimated water table of 0–500 m, 500–1250 m, and 1250–2750 m. The model was calibrated using the 14 guidelines in Table 1 and the inverse groundwater flow model MODFLOWP (Hill, 1992). Application of guidelines 1 to 10 and parts of 12 and 13 are described briefly in Table 1. This paper focuses on using the model developed thus far to conduct the analyses suggested by guidelines 12 and 14.

GUIDELINE 12: EVALUATE THE POTENTIAL FOR ADDITIONAL ESTIMATED PARAMETERS

Composite scaled sensitivities and parameter correlation coefficients (Hill, 1998, p. 15 and 28) can be used to measure the amount of information provided by the regression

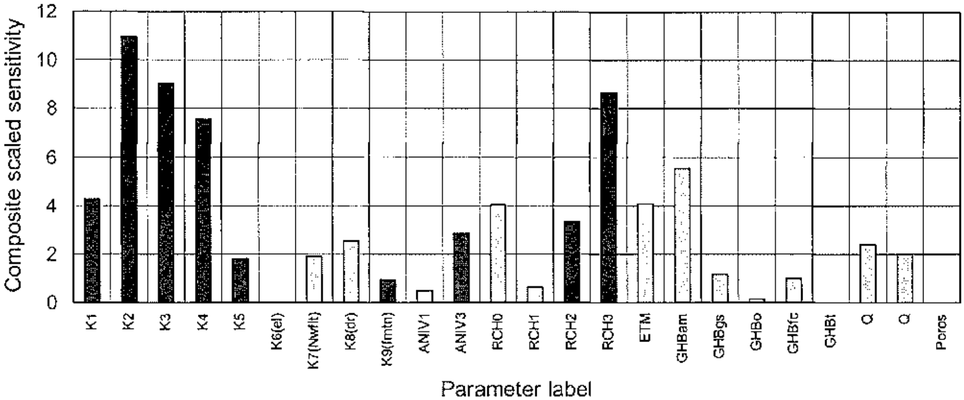

Fig. 1 Composite scaled sensitivities for all parameters defined for the final Death Valley Regional groundwater flow system model. The bars are black for parameters estimated by regression; grey otherwise. Parameters are log transformed as they were for the regression. (Parameter labels: *K*, hydraulic conductivity; *AVIV*, vertical anisotropy; *RCH*, areal recharge; *ETM*, maximum evapotranspiration; *GHB*, spring conductance; *Q*, pumping rate; *Poros*, porosity).

data (the measurements used as observations in the regression) with which to estimate defined parameters uniquely. The composite scaled sensitivities for all 23 defined parameters of the final Death Valley regional model, calculated at the optimized parameter values, are shown in Fig. 1. Porosity is included in the figure to coordinate with later figures; its composite scaled sensitivity equals 0.0 because it has no affect on simulated hydraulic heads and volumetric flows such as spring flows.

Composite scaled sensitivities for each parameter b_j are calculated as $[1/n\sum((\partial y'_i/\partial b_j)b_j\omega_i^{1/2})^2]^{1/2}$ where y'_i is one of the n observations, the summation is over i, and ω_i is a weight. Composite scaled sensitivities reflect the amount of information the regression data contain about each parameter. In the absence of extreme parameter correlation, parts of the system represented by parameters with large composite scaled sensitivities usually can be represented with more parameters. The large composite scaled sensitivity for *K*3, for instance, indicates that the parts of the system represented by the *K*3 parameter (which includes some types of metamorphic rocks, basalt, tuff, and basin fill) can be subdivided to represent some or all of these units independently. Parameter *K*6, which represents the shale-rich Eleana Formation and surrounding rocks coincident with a large hydraulic gradient north and east of Yucca Mountain, has a small composite scaled sensitivity which indicates that, although the existence of low hydraulic conductivity there is important, the simulated value is so small that the hydraulic head and spring-flow regression data are insensitive to large fractional variations about the simulated value.

Without the spring-flow data, only hydraulic head data would be available for calibration. When all flows and hydraulic properties are defined as parameters, as in the Death Valley regional model, using only hydraulic head observations in the regression results in all parameters being extremely correlated (parameter correlation coefficients of 1.0 or −1.0) and no parameter values could be uniquely estimated. With the spring-flow data, the parameter correlations are mostly less than 0.95, with exceptions shown in Table 2. Thus, the available spring-flow data are sufficient to

Table 2 Parameter correlations with absolute values greater than 0.95.

Parameter labels		Correlation	Parameter labels		Correlation
*GHB*gs	*K*7(Nwflt)	–0.99	*RCH*3	*Q*2	0.97
*K*2	*RCH*3	0.98	*K*1	*RCH*3	0.96

allow lost parameter values to be estimated uniquely. The largest correlation is between *GHB*gs, the spring conductance of Grapevine Springs, and *K*7(Nwflt), the hydraulic conductivity of a set of northwest trending faults. The large correlation of the *K*7(Nwflt) and *GHB*gs parameters and the associated high sensitivity of these parameters to observations of spring flow at Grapevine Springs (from a sensitivity analysis not presented in this work), indicate that neither of these parameters probably influences much relevant to the observed hydraulic heads and flows except at Grapevine Springs, where they largely fulfil the same role. The remaining large correlations involve flows, indicating that additional and more precisely measured flow observations are likely to improve future models. The effects of such improved measurements and additional data can be evaluated as part of guideline 11 (Hill, 1998, p. 55–57). Such an analysis is being conducted, but is beyond the scope of this paper.

GUIDELINE 14: FORMALLY RECONSIDER THE MODEL CALIBRATION FROM THE PERSPECTIVE OF THE DESIRED PREDICTIONS

The predictions of most interest in the Death Valley regional model involve potential transport of contaminants from beneath Yucca Mountain and from nuclear testing sites. Simulation of transport is complicated by the scale of the regional model, which makes it difficult to accurately represent the smaller-scale features that affect transport. Thus, a useful and necessary approach is to consider only some of the transport processes. Advective transport is chosen for consideration because it is the aspect of transport most affected by regional flow characteristics. By definition, advective transport occurs when solute does not spread through dispersive processes and does not react with the surrounding rocks. Advective transport is produced, on average, by bulk flow in the subsurface, and can be thought of as an initial calculation of transport that can be built upon by adding complexities.

In this work, advective transport is calculated as the distance and path travelled over a specified length of time. The path travelled is subdivided into the three spatial model directions: for this model, north–south, east–west, and vertical. This formulation is consistent with that of the ADV Package (Anderman & Hill, 1997) to MODFLOWP, which was used to simulate advective transport in this work. The ADV Package calculates advective transport using particle tracking implemented as in MODPATH (Pollock, 1989).

Evaluation of very long advective travel paths is avoided to minimize the discrepancies between the simulated and actual travel paths, which are likely to be larger for longer flow paths. This approach also is consistent with the importance of accurately representing the transport characteristics of flow paths near the source of underground testing areas and Yucca Mountain. As noted below, sensitivities are presented for an advective travel distance of 10 000 m, or about 7 grid-dimension lengths.

The relation of the simulated time to the actual time of advective transport depends on effective porosity. Widely varying porosity values are attributed to the rocks and sediments (values of 0.0001 to 0.37 are cited by Bedinger *et al.*, 1989). However, a uniform porosity of 0.10, is used in the simulations. This means that the simulated times are considerably less reliable than the simulated advective travel paths. To emphasize results that are more likely to be accurate, times are presented as scaled by a reference time and velocities are presented relative to the maximum velocity calculated along the travel path. The sensitivities are likely to be affected by the assumed constant value of porosity in that they are likely to indicate diminished or enhanced apparent importance of rocks with extreme porosity values.

The results presented here are for transport from one of the underground testing sites. The site is located within one finite-difference cell; advective transport from the water table at the cell centre and near the four cell corners was tracked to capture the effect of small changes in origin. The relative velocity and time of advective travel of transport from the cell centre are shown in Fig. 2, and indicate that the largest velocities occur between 57 000 and 78 000 m along the travel path.

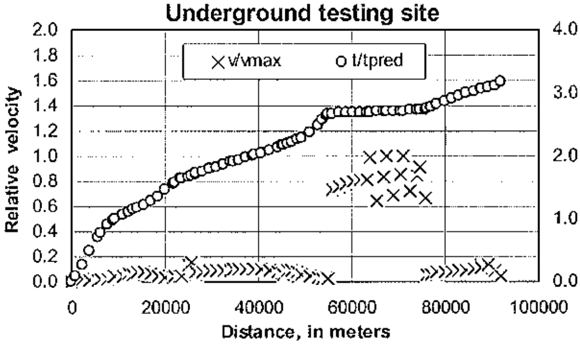

Fig. 2 Relative velocity and time since release of particle started at the water table at the cell centre at one of the underground testing areas. Relative velocity equals velocity divided by the maximum velocity (*vmax*) along the particle path. Relative time equals the time divided by the time (*tpred*) required for the particle to move 10 000 m, which is the distance for which the sensitivities of Fig. 3 are calculated. The large velocities between travel distances of 57 000 and 78 000 m are caused by local variations in subsurface conditions.

Prediction scaled sensitivities are calculated as $|(\partial z_\ell/\partial b_j)(b_j/100)(1/z_\ell) \times 100|$, where z_ℓ is the ℓth prediction, b_j is the jth parameter (not log transformed), and multiplication by 100 produces a percent instead of a fraction of the simulated prediction. Prediction scaled sensitivities approximately equal the percent change in simulated travel distance in the applicable direction produced by a one percent increase in the parameter value. For the recharge parameters, the prediction scaled sensitivity is the amount the prediction would change if the recharge increased by one percent of local annual average precipitation. The heavy horizontal lines in the graphs of Fig. 3 mark where a one percent change in the parameter value produces a one percent change in the simulated value. The smallest plotted value of 0.0001 indicates that the parameter value would need to change by a factor of 100 (10 000%) to produce a one percent

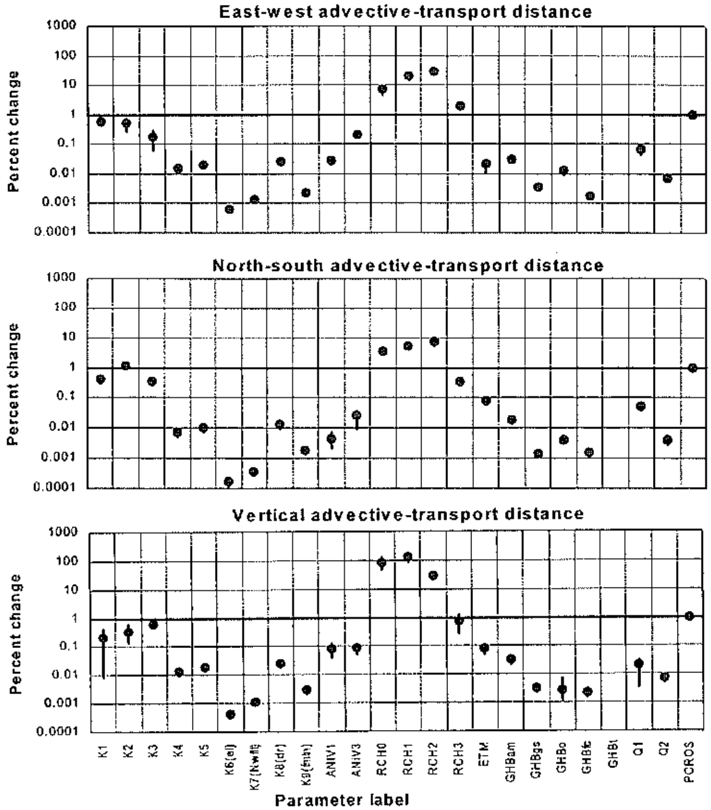

Fig. 3 Prediction scaled sensitivities for the testing area of Fig. 2 defined to approximate the absolute value of the percent that advective transport in the direction indicated would change given a one-percent change in the parameter value. The range for five particles is shown by a vertical line.

change in the simulated value. Absence of a symbol for *GHBt* means that all values were less than 0.0001. Sensitivities are calculated for a total particle travel distance of 10 000 m (to 5250 m west, 6853 m south, and 128 m down from the starting point; distance is based on actual path, not a straight line from the starting point). The prediction scaled sensitivity for the global porosity parameter equals 1.0 because a one percent change in all porosities (even if they are not uniform in value) changes advective-travel in all directions by one percent.

Prediction scaled sensitivities will change along the flow path. The robustness of the conclusions drawn from the sensitivity analysis depends on the changes being moderate. Prediction scaled sensitivities for the east–west, north–south, and vertical directions are shown in Fig. 4 for parameter *RCH0*, which, at a 10 000 m travel distance, has a large prediction scaled sensitivity (Fig. 3). For this parameter, the prediction scaled sensitivities are stable for about 50 000 m of transport distance.

The composite scaled sensitivities and the prediction scaled sensitivities can be used to identify parameters most likely to contribute to prediction uncertainty. The procedure is outlined briefly in Fig. 5, and in detail by Hill (1998, p. 62–64). Confidence intervals and parameter correlation coefficients also can, and should, be

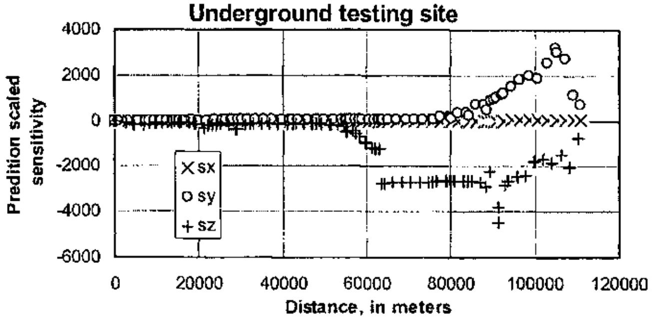

Fig. 4 Variation of the prediction scaled sensitivities for parameter RCH0 for advective travel in the x (east–west) and y (north–south) directions with distance travelled from the source.

		Likely precision of the parameter estimate	
		Poor: Small composite scaled sensitivity	**Good:** Large composite scaled sensitivity
Importance of the parameter to predictions of interest	**Not important:** Small prediction scaled sensitivity	I. Acceptable	II. Acceptable
	Important: Large prediction scaled sensitivity	IV. Improve estimation of this parameter and representation of associated system features	III. Acceptable

Fig. 5 Using composite and prediction scaled sensitivities to identify critical parameters.

similarly evaluated but, for brevity, are not presented here. Analysis of composite and prediction scaled sensitivities indicate that for many parameters important to advective travel, the regression data provide substantial information, which is good. Of most concern is *RCH*1 because the regression data do not provide much information about this parameter (Fig. 1), but *RCH*1 is important to predicted advective travel (Fig. 3).

SUMMARY AND CONCLUSIONS

The primary advantages of the model calibration approach presented in this paper are:

- it is more objective and tests the model against measured data more rigorously than most alternatives, and
- the statistics are designed to communicate the strengths and weaknesses of the simulations clearly to non-modellers such as resource managers, funding agencies, and interested citizens.

The analysis of the Death Valley regional flow system presented indicates that the available hydraulic head and spring-flow data used in the regression provide substantial information relevant to the advective transport predictions. There appears to be sufficient information in the existing regression data to estimate many aspects of the hydraulic-conductivity field and areal recharge distribution with additional defined parameters, and the achievable detail will be greatly enhanced if additional and more precise measurements of flows into, out of, and through the system can be attained to reduce parameter correlation.

REFERENCES

Anderman, E. R. & Hill, M. C. (1997) Advective transport observations (ADV) Package, A computer program for adding advective transport observations of steady-state flow fields to the three-dimensional ground-water flow parameter-estimation model MODFLOWP. *US Geological Survey Open-File Report 97-14.*

Barth, G. R., Hill, M. C., Illangasekare, T. H. & Rajaram, Harihar (1999) Predictive modeling of flow in a two-dimensional intermediate-scale, heterogeneous porous media. In: *Calibration and Reliability in Groundwater Modelling* (Proc. ModelCARE 99 Conference held at Zürich, Switzerland, September 1999). IAHS Publ. no. 265. This volume.

Bedinger, M. S., Langer, W. H., & Reed, J. E. (1989) Hydraulic properties of rocks in the Basin and Range Province. In: *Studies of Geology and Hydrology in the Basin and Range Province, Southwestern United States, for Isolation of High-level Radioactive Waste: Basis of Characterization and Evaluation* (ed. by M. S. Bedinger, K. A. Sargent, W. H. Langer, F. B. Sherman, J. E. Reed & B. T. Brady). US Geological Survey Professional Paper 1370-A.

D'Agnese, F. A., Faunt, C. C., Turner, A. K. & Hill, M. C. (1998) Hydrogeologic evaluation and numerical simulation of the Death Valley regional ground-water system, Nevada and California, using geoscientific information systems. *US Geological Survey Water-Resources Investigations Report 4300.*

D'Agnese, F. A., Faunt, C. C., Hill, M. C. & Turner, A. K. (1999) Death Valley regional ground-water flow model calibration using optimal parameter estimation methods and geoscientific information systems. *Adv. Wat. Resour.* 22(8), 777–790.

Doherty, J. (1994) *PEST.* Watermark Computing, Corinda, Australia.

Hill, M. C. (1992) A computer program (MODFLOWP) for estimating parameters of a transient, three-dimensional ground-water flow model using nonlinear regression. *US Geological Survey Open-File Report 91-484.*

Hill, M. C. (1998) Methods and guidelines for effective model calibration. *US Geological Survey Water-Resources Investigations Report 98-4005.*

Kitanidis, P. K. (1998) How observations and structure affect the geostatistical solution to the steady-state inverse problem. *Ground Water* 35(5), 754–763.

Mehl, S., Hill, M. C. & Illangesakare, T. (1999) A comparison of solute-transport solution techniques based on inverse modelling results. In: *Calibration and Reliability in Groundwater Modelling* (Proc. ModelCARE 99 Conference held at Zürich, Switzerland, September 1999). IAHS Publ. no. 265. This volume.

Poeter, E. P. & Hill, M. C. (1997) Inverse modeling: A necessary next step in ground-water modeling. *Ground Water* 35(2), 250–260.

Poeter, E. P. & Hill, M. C. (1998) Documentation of UCODE, a computer code for universal inverse modeling. *US Geological Survey Water-Resources Investigations Report 98-4080.*

Pollock, D. W. (1989) Documentation of computer programs to compute and display pathlines using results from the US Geological Survey modular three-dimensional finite-difference ground-water flow model. *US Geological Survey Open-File Report 89-381.*